CoVLM: Leveraging Consensus from Vision-Language Models for Semi-supervised Multi-modal Fake News Detection

Devank, Jayateja Kalla, and Soma Biswas

Department of Electrical Engineering,
Indian Institute of Science, Bangalore, India.
{devank2022, jayatejak, somabiswas}@iisc.ac.in

Abstract. In this work, we address the real-world, challenging task of out-of-context misinformation detection, where a real image is paired with an incorrect caption for creating fake news. Existing approaches for this task assume the availability of large amounts of labeled data, which is often impractical in real-world, since it requires extensive manual intervention and domain expertise. In contrast, since obtaining a large corpus of unlabeled image-text pairs is much easier, here, we propose a semi-supervised protocol, where the model has access to a limited number of labeled image-text pairs and a large corpus of unlabeled pairs. Additionally, the occurrence of fake news being much lesser compared to the real ones, the datasets tend to be highly imbalanced, thus making the task even more challenging. Towards this goal, we propose a novel framework, **Co**nsensus from **V**ision-**L**anguage **M**odels (CoVLM) [1], which generates robust pseudo-labels for unlabeled pairs using thresholds derived from the labeled data. This approach can automatically determine the right threshold parameters of the model for selecting the confident pseudo-labels. Experimental results on benchmark datasets across challenging conditions and comparisons with state-of-the-art approaches demonstrate the effectiveness of our framework.

Keywords: Vision-Language Models, Semi-Supervised Learning, multi-modal Fake News Detection

1 Introduction

The proliferation of fake news in social media has made fake news detection a critical task for maintaining information integrity [1], safeguarding public discourse [2], and preventing the erosion of trust [3]. One of the increasingly popular means of generating fake news is to pair real images with misleading/incorrect captions, since this requires minimal effort and technical expertise. Figure 1 demonstrates few examples of real and fake image-text pairs from the benchmark NewsCLIPings Dataset [4]. Fake news often exhibit discrepancies between

[1] Code Link https://github.com/devank3/CoVLM

Fig. 1: A sample real and fake image-pair from the NewsCLIPpings dataset [4]. The model needs to capture the subtle inconsistencies between the image and text pairs to understand their authenticity.

the visual content and the accompanying text, whereas real news tends to have a coherent relationship between images and text. Identifying these subtle inconsistencies can help to determine whether a given image-text pair is real or fake. Thus the goal of the existing out-of-context misinformation detection or multimodal fake news detection (MFND) frameworks is to analyze large amounts of training data to learn these inconsistencies, which are used to infer whether a given test image-text pair is real or fake.

One of the significant advancements in the direction of multi-modal machine learning [5–7] is to learn joint representations of image content and natural language. For example, the CLIP (Contrastive Language-Image Pre-training) model [8] bridges the gap between image and natural language modalities by being trained on a huge dataset of image-caption pairs. Approaches leveraging these models for MFND task [4, 9–15] have shown promise. However, these approaches rely entirely on supervised data, i.e., image-text pairs labeled as real or fake. Annotating large amounts of data is extremely labor-intensive and requires domain expertise. For instance, verifying a news claim like *"An anti-government protester waves a Thai national flag outside Parliament in Bangkok"* as demonstrated in Figure 1 involves significant expertise in international affairs and requires professionals who are proficient in global political dynamics. Conversely, collecting image-text pairs without annotations is much simpler. In this work, we propose a realistic and practical Semi-Supervised MFND (SS-MFND) protocol, where the model has access to a few labeled image-text pairs and a large number of unlabeled pairs.

One of the most successful approaches to leverage unlabeled data in semi-supervised learning [16] involves training a model first on the labeled data, followed by generating pseudo-labels for the unlabeled data and incorporating the confident ones into the training process. While such methods, like FixMatch [16], Adsh [17], and FreeMatch [18], have proven effective for unimodal data by utilizing the vast amounts of available unlabeled data, they fall short in semi-supervised MFND because they fail to capture the intricate relationships between real and fake image-text pairs. To address this challenge and generate robust pseudo-labels for unlabeled image-text pairs, we propose CoVLM, a

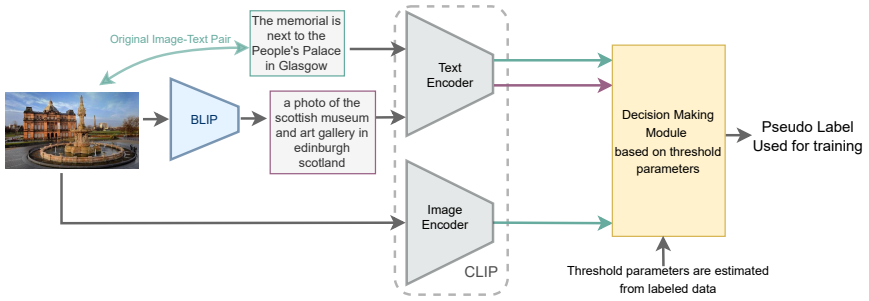


Fig. 2: Overview of CoVLM. For a given image-text pair, BLIP generates an additional image description. Using the original image, text, and the generated text, a decision is made on whether the pair is real or fake. This label is then used in training. The decision module's threshold parameters are estimated from the labeled data.

novel approach that leverages the consensus between two vision-language models (VLMs), namely CLIP (Contrastive Language-Image Pre-training) [8] and BLIP (Bootstrapping Language-Image Pre-training) [19]. The BLIP model generates descriptive text for a given image, which is then used in conjunction with the original text to generate robust pseudo-labels. This consensus from two models (CLIP, BLIP) ensures the robustness of the pseudo-labels. Figure 2 provides an overview of the CoVLM using CLIP and BLIP models.

Another challenge is that for image-text pairs collected from news-articles, blogs, posts, etc. the number of real pairs are far more compared to the number of fake pairs. This results in severely class-imbalanced data, thereby making the task significantly more challenging. Since the benchmark NewsCLIPpings data is artificially created and balanced, it is not suited to analyze the performance of algorithms under a more realistic imbalanced scenario. Inspired by the rich literature on imbalanced semi-supervised learning for classification tasks [17, 20–27], we introduce such imbalances in labeled and unlabeled data to evaluate the proposed approach under these realistic conditions. To this end, the contributions of this work can be summarized as follows:

1. To the best of our knowledge, this is the first work to address the realistic and challenging Semi-Supervised multi-modal Fake News Detection task.
2. We propose a novel framework CoVLM that utilizes vision-language models to generate robust pseudo-labels for the unlabeled image-text training pairs.
3. Extensive experiments on widely used MFND datasets, namely NewsCLIPpings [4], GossipCop [28], and PolitiFact [29], demonstrate the effectiveness of the proposed approach.
4. In addition to the traditional balanced settings, we also test the framework in realistic imbalanced scenarios to evaluate its robustness.

We now discuss the related work followed by the proposed framework and evaluation results.

2 Related Work

In this section we briefly discuss the related works on fake news detection and semi-supervised learning.

Unimodal Fake News Detection: Traditional approaches to fake news detection often focus on analyzing a single modality, such as text or image content in isolation. (i) Image Analysis: Prior research has explored image forensic features, semantic information, and statistical properties to detect manipulation [30]. Techniques for identifying image tampering can reveal signs of fake [31]. Additionally, common sense inconsistencies and poor image quality can be red flags for fake news [32, 33]. (ii) Text Analysis: Verifying logical consistency is crucial for detecting fake news in text format [34]. Examining for grammatical errors, unusual writing styles, or specific rhetorical structures can also provide clues [34, 35]. However, both linguistic and visual patterns can be heavily influenced by the specific event and related domain knowledge. To address this challenge, Nan et al. [36] proposed utilizing a domain gate to combine representations learned from different experts, enabling their model to handle multi-domain fake news propagation within the text modality. While these unimodal features offer valuable insights and play a significant role in distinguishing fake news, they neglect the crucial aspects of multi-modality - correlation and consistency between text and image content. This omission can hinder the overall effectiveness of these single modality methods when applied to multi-modal news.

Multimodal Fake News Detection: MFND focuses on utilizing both text and image modalities to detect fake news. Several approaches have been proposed to address this task, few notable ones being SAFE [37], Cultural Algorithm [38], Combination of textual, visual, and semantic information [39], Fine-grained Classification [40], FND-CLIP [14], ETMA [41], SAMPLE [15], DeBERTNeXT [42] and Tri Transformer-BLIP [43]. Recently, researchers have also started gathering evidence from the internet to determine the authenticity of the image-text pair, and some notable works in this direction include NewsCLIPpings [4], Consistency-Checking Network [9], SEN [10], OOCD [11], SNIFFER [12], Zero-shot approach [44], EVVER-Net [45], and MMFakeBench [13]. But all these approaches assume that the entire training data is labeled, which is quite difficult for this task in realistic scenarios, because of the amount of human intervention required. Thus, we propose a real-world semi-supervised setting, which can utilize unlabeled data along with some labeled pairs for addressing this task. We also work in the closed setting, with no external evidence.

Semi-Supervised Learning: In semi-supervised learning (SSL), various methods have been developed to effectively leverage unlabeled unimodal data to improve model performance. Numerous approaches have been proposed in the literature for the image domain [16, 18, 46–55]. FixMatch [16] is a significant work in this field, combining consistency regularization and pseudo-labeling with fixed thresholds. Building on the ideas of FixMatch, Adsh [17] proposed using dif-

ferent thresholds for different classes to help underrepresented classes improve accuracy. SSL in the text domain also has been extensively explored and continues to evolve with new methodologies and approaches [56–63]. These works primarily focus on individual modalities either in the image or text domains within the context of SSL. However, our research aims to bridge these methodologies by concentrating on fake news detection involving both image and text components in a semi-supervised manner.

3 Problem Definition

We now formally define the problem of semi-supervised multi-modal fake news detection (SS-MFND), where the objective is to train a model using a limited amount of labeled data and large amount of unlabeled training data, to determine whether a given image-text pair is real or fake at inference time. Specifically, the model has access to a labeled data $\mathcal{D}^l = \{(\mathcal{I}_i^l, \mathcal{T}_i^l, y_i^l)\}_{i=1}^{N^l}$ which consists of image-text pairs $(\mathcal{I}_i^l, \mathcal{T}_i^l)$ along with their corresponding labels y_i^l, indicating whether the pair is real or fake. Here, N^l denotes the number of labeled data samples. Additionally, the model has access to an unlabeled dataset $\mathcal{D}^{ul} = \{(\mathcal{I}_i^{ul}, \mathcal{T}_i^{ul})\}_{i=1}^{N^{ul}}$, which contains image-text pairs without any information regarding their authenticity. It is usually easier to collect the image-text pairs compared to labeling them, which requires significant manual intervention and domain expertise. Notably, $N^l \ll N^{ul}$, indicating that the number of labeled samples is much smaller than the number of unlabeled samples in our experiments. We now discuss the proposed CoVLM framework in detail.

4 CoVLM for Semi-Supervised MFND

In our work, inspired by the recent literature in MFND [4, 9–11, 14], we utilize the powerful CLIP model because of its shared latent space, learnt from image-caption pairs. To effectively utilize the unlabeled data, we want to generate robust pseudo-labels which can then be used for further training the model. The core idea of CoVLM is to leverage additional guidance from another vision-language model BLIP, which can convert the input image into descriptive text. This generated text, combined with the original image-text pair, is used to create robust pseudo-labels based on threshold parameters derived from the labeled data. The intuition behind determining the pseudo-labels for the unlabeled data using threshold parameters is detailed in Subsection 4.1. Subsection 4.2 explains how the threshold parameters are obtained from the labeled data. Finally, Subsection 4.3 outlines the complete training procedure, incorporating both labeled and unlabeled data using the pseudo-labels.

4.1 Unlabeled data: Pseudo-Labels using Caption Consensus

Here, we explain how we determine robust pseudo-labels for the unlabeled data in the training set. Solely using the CLIP model to generate real/fake pseudo-labels is challenging, since realistic fake image-text pairs are semantically very

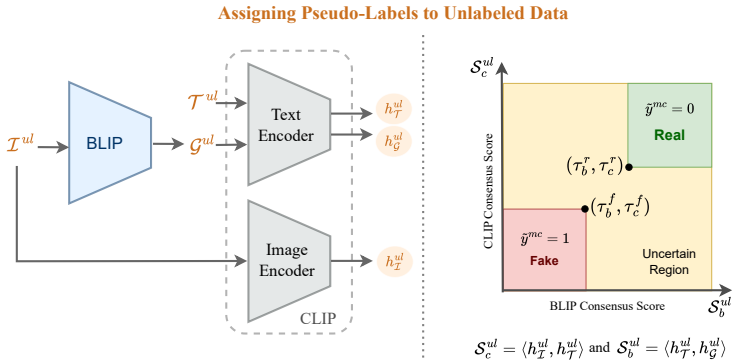


Fig. 3: Illustration of the pseudo-label assignment for the unlabeled image-text pairs. Using both image and text embeddings, CLIP consensus score $\mathcal{S}_c$ is calculated and for given text and generated text BLIP consensus score $\mathcal{S}_b$ is calculated.

close, and it is difficult to differentiate them from actual real pairs. Towards this goal, we propose to use consensus from two vision-language models (VLMs), namely CLIP and BLIP. For simplicity, we denote the CLIP model as Θ^{CLIP}, which includes both the image and text encoders represented as $\{\Theta^{\text{CLIP}}_{\text{image}}, \Theta^{\text{CLIP}}_{\text{text}}\}$ and the BLIP model as Θ^{BLIP}. First we pass each unlabeled image $\mathcal{I}^{ul}$ through BLIP to obtain a generated caption denoted as $\mathcal{G}^{ul} = \Theta^{\text{BLIP}}(\mathcal{I}^{ul})$. The image and text embeddings of this unlabeled image-text pair obtained from the CLIP model is denoted as $\{h^{ul}_{\mathcal{I}}, h^{ul}_{\mathcal{T}}\}$, by passing the image $\mathcal{I}^{ul}$ and text $\mathcal{T}^{ul}$ through the image and text encoders respectively. We also compute the embedding of the BLIP-generated text caption as $h^{ul}_{\mathcal{G}} = \Theta^{\text{CLIP}}_{\text{text}}(\mathcal{G}^{ul})$. Now, for a real image-text pair, their corresponding embeddings in the CLIP shared latent space will be relatively closer compared to when the pair is fake. This is measured by the CLIP consensus score $\mathcal{S}^{ul}_c = \langle h^{ul}_{\mathcal{I}}, h^{ul}_{\mathcal{T}} \rangle$, where $\langle \cdot, \cdot \rangle$ represents the inner product between two vectors, and measures the similarity between them. Similarly, the BLIP consensus score, which is calculated between the embeddings of the generated text from BLIP and that of the original text as $\mathcal{S}^{ul}_b = \langle h^{ul}_{\mathcal{T}}, h^{ul}_{\mathcal{G}} \rangle$ is high for real pairs and low for fake pairs. Thus, for a real image-text pair, both the scores will be high, whereas, for fake pairs, both scores will be low, as the image and text will disagree, in addition to the original and generated text being far apart. This model consensus ensures generation of robust pseudo-labels for the unlabeled data. Unlabeled samples for which the models are not in agreement are not used for training at that instant. Note that at a later training instance, this particular image-text pair can be assigned a confident pseudo-label and can thus contribute to model training. The complete multi-modal consensus generated pseudo-labels

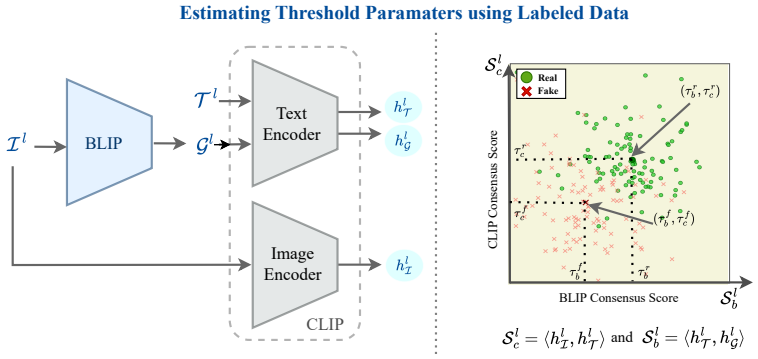


Fig. 4: Illustration of the estimation of threshold parameters using labeled data. The BLIP consensus score $\mathcal{S}_b$ and the CLIP consensus score $\mathcal{S}_b$ are calculated for all labeled samples. The mean of these labeled BLIP and CLIP consensus scores acts as threshold parameters for the unlabeled data.

is given as:

$$\tilde{y}^{mc} = \begin{cases} Fake & \text{if } \mathcal{S}_c^{ul} < \tau_c^f \text{ and } \mathcal{S}_b^{ul} < \tau_b^f \\ Real & \text{if } \mathcal{S}_c^{ul} > \tau_c^r \text{ and } \mathcal{S}_b^{ul} > \tau_b^r \\ Ignore & \text{Otherwise} \end{cases} \tag{1}$$

The threshold boundary parameters for fake samples $\{\tau_c^f, \tau_b^f\}$ and real samples $\{\tau_c^r, \tau_b^r\}$ play a key role in the decision-making process of pseudo-labels and are dataset dependent. We utilize the labeled part of the training dataset to automatically obtain these parameters. Figure 3 shows the pseudo label assignment to the image-text pairs based the model consensus using CLIP and BLIP models. Next, we discuss how these boundary thresholds are estimated.

4.2 Threshold Parameters from Labeled Data

We utilize the available labeled data to automatically determine the threshold parameters, which removes the burden of manually tuning the parameters for each dataset. For a labeled image-text pair $(\mathcal{I}^l, \mathcal{T}^l, y^l)$, let the BLIP generated caption be denoted as $\mathcal{G}^l = \Theta^{\text{BLIP}}(\mathcal{I}^l)$. Now, the CLIP embeddings for the given image and text pair are given by $h_{\mathcal{I}}^l = \Theta_{\text{image}}^{\text{CLIP}}(\mathcal{I}^l)$ and $h_{\mathcal{T}}^l = \Theta_{\text{text}}^{\text{CLIP}}(\mathcal{T}^l)$, and the embedding of the BLIP-generated text caption is given by $h_{\mathcal{G}}^l = \Theta_{\text{text}}^{\text{CLIP}}(\mathcal{G}^l)$. As shown in Figure 4, we calculate the consensus scores of BLIP model, $\mathcal{S}_b^l = \langle h_{\mathcal{T}}^l, h_{\mathcal{G}}^l \rangle$, and for the CLIP model, $\mathcal{S}_c^l = \langle h_{\mathcal{I}}^l, h_{\mathcal{T}}^l \rangle$ for all the labeled samples. Using the label y^l, we calculate the mean of these scores for both real and fake samples to obtain these threshold parameters. Specifically, the threshold parameters for the real class are calculated as $\tau_b^r = \sum_i \mathcal{S}_b^i \cdot \mathbb{I}(y_i^l = Real)$, $\tau_c^r = \sum_i \mathcal{S}_c^i \cdot \mathbb{I}(y_i^l = Real)$ (the superscript l for model score $\mathcal{S}$ is removed

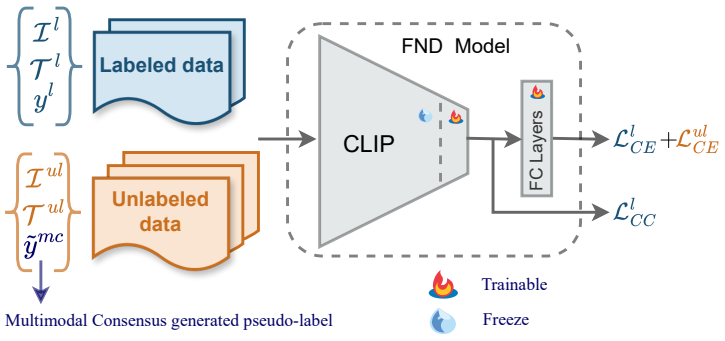


Fig. 5: Illustration of the unified training process. Both labeled and unlabeled data are used for training with binary cross-entropy loss. Contrastive clustering loss, applied to labeled data, ensures that real image-text pair embeddings are close together, while forcing fake pair embeddings far apart. This complements the pseudo-label generation process and enhances the performance of semi-supervised MFND.

for clarity). In a similar manner, we calculate the parameters for the fake class. Here, $\mathbb{I}(\cdot)$ denotes the indicator function, which is 1 if the condition is true and 0 otherwise. After obtaining pseudo-labels for the unlabeled data using these parameters, the model training proceeds using both the labeled and unlabeled data, as discussed in the next section. Figure 5 illustrates the unified training process of the CoVLM framework.

4.3 Unified Training using both Labeled and Unlabeled Data

Now, we describe the complete training process of the proposed framework for semi-supervised MFND using both the labeled and the unlabeled inputs. During training, the last few layers of CLIP image encoder $\Theta^{\text{CLIP}}_{\text{image}}$ along with two fully connected layer ψ are learned. The predicted output for an image-text pair $(\mathcal{I}, \mathcal{T})$ is $\hat{y} = \psi(\Theta^{\text{CLIP}}_{\text{image}}(\mathcal{I}) \odot \Theta^{\text{CLIP}}_{\text{text}}(\mathcal{T}))$, where $\odot$ represents the hadamard product (element wise multiplication).

Learning from Labeled Data: For the labeled dataset $\mathcal{D}^l$, the cross-entropy loss is computed between the model's predictions and the true labels. Let $\hat{y}_i$ be the model's predicted output for the i^{th} image-caption pair $(\mathcal{I}^l_i, \mathcal{T}^l_i)$, and y^l_i be the true label. The binary cross-entropy loss $\mathcal{L}^l_{CE}$ is calculated by:

$$\mathcal{L}^l_{CE} = -\frac{1}{N_l}\sum_{i=1}^{N_l}\left(y^l_i \log \hat{y}_i + (1-y^l_i)\log(1-\hat{y}_i)\right) \tag{2}$$

Our core assumption in addressing the MFND task is that real image-text pairs will be closer in the embedding space, while fake pairs will be far apart and the

Algorithm 1 Training Algorithm

Input Model: $\Theta^{\text{CLIP}} = (\Theta^{\text{CLIP}}_{\text{image}}, \Theta^{\text{CLIP}}_{\text{text}})$, $\Theta^{\text{BLIP}}, \psi -$ FC Layers
Training Data: $\mathcal{D}^l = \{(\mathcal{I}^l_i, \mathcal{T}^l_i, y^l_i)\}^{N^l}_{i=1}$, $\mathcal{D}^{ul} = \{(\mathcal{I}^{ul}_i, \mathcal{T}^{ul}_i)\}^{N^{ul}}_{i=1}$ where $N^l \ll N^{ul}$
for each *epoch* **do**
  $\mathcal{B}^l$ = SampleMiniBatch($\mathcal{D}^l$) and B^{ul} = SampleMiniBatch($\mathcal{D}^{ul}$)
  // Generating BLIP captions and CLIP embeddings //
  $\mathcal{G}^l \leftarrow \Theta^{\text{BLIP}}(\mathcal{B}^l)$; $\mathcal{G}^{ul} \leftarrow \Theta^{\text{BLIP}}(\mathcal{B}^{ul})$
  $h^l_{\mathcal{I}}, h^l_{\mathcal{T}}, h^l_{\mathcal{G}} \leftarrow \Theta^{\text{CLIP}}(\mathcal{B}^l, \mathcal{G}^l)$
  $h^{ul}_{\mathcal{I}}, h^{ul}_{\mathcal{T}}, h^{ul}_{\mathcal{G}} \leftarrow \Theta^{\text{CLIP}}(\mathcal{B}^{ul}, \mathcal{G}^{ul})$
  // Estimating threshold parameters from labeled data //
  $\mathcal{S}^l_c = \langle h^l_{\mathcal{I}}, h^l_{\mathcal{T}} \rangle$; $\mathcal{S}^l_b = \langle h^l_{\mathcal{I}}, h^l_{\mathcal{G}} \rangle$
  $\tau^f_c, \tau^r_c, \tau^f_b, \tau^r_b \leftarrow$ Estimate threshold parameters using $\mathcal{S}^l_c$, $\mathcal{S}^l_b$ (Subsection. 4.2)
  // Assigning pseudo-labels to unlabeled data //
  $\mathcal{S}^{ul}_c = \langle h^{ul}_{\mathcal{I}}, h^{ul}_{\mathcal{T}} \rangle$; $\mathcal{S}^{ul}_b = \langle h^{ul}_{\mathcal{I}}, h^{ul}_{\mathcal{G}} \rangle$
  $\tilde{y}^{\text{mc}} \leftarrow$ Assign pseudo label using $\mathcal{S}^{ul}_c$, $\mathcal{S}^{ul}_b$, and estimated thresholds (Eq. 1)
  // Output model predictions and loss calculations //
  $\hat{y}^l \leftarrow \psi(h^l_{\mathcal{I}} \odot h^l_{\mathcal{T}})$; $\hat{y}^{ul} \leftarrow \psi(h^{ul}_{\mathcal{I}} \odot h^{ul}_{\mathcal{T}})$
  $\mathcal{L}^l_{CE} = \text{CrossEntropyLoss}(y^l, \hat{y}^l)$, $\mathcal{L}^l_{CC} = \text{ContrastiveClusterLoss}(y^l, \mathcal{S}^l_c)$
  $\mathcal{L}^{ul}_{CE} = \text{CrossEntropyLoss}(\tilde{y}^{\text{mc}}, \hat{y}^{ul})$
  $\mathcal{L}_{Total} = \mathcal{L}^l_{CE} + \mathcal{L}^{ul}_{CE} + \lambda \mathcal{L}^l_{CC}$
end for
return $\{\Theta^{\text{CLIP}}, \psi\}$

Algorithm 2 Inference Algorithm

Trained CLIP Model: $\Theta^{\text{CLIP}} = (\Theta^{\text{CLIP}}_{\text{image}}, \Theta^{\text{CLIP}}_{\text{text}}), \psi -$ FC Layers
Test data: $\mathcal{D}^t = (\mathcal{I}^t_i, \mathcal{T}^t_i)^{N^t}_{i=1}$
for each $i \in 1, \ldots, N^t$ **do**
  $\hat{y}_i \leftarrow \psi(\Theta^{\text{CLIP}}_{\text{Image}}(\mathcal{I}^t_i) \odot \Theta^{\text{CLIP}}_{\text{Text}}(\mathcal{T}^t_i))$
end for

pseudo-labeling for unlabeled image-text pairs also relies on this principle. To enforce this notion within the network, we propose to use an additional clustering objective inspired by contrastive loss given as:

$$\mathcal{L}^l_{CC} = -\frac{1}{N_l} \sum_{i=1}^{N_l} \left(y^l_i \log(1 - \langle h^l_{\mathcal{I}}, h^l_{\mathcal{T}} \rangle) + (1 - y^l_i) \log \langle h^l_{\mathcal{I}}, h^l_{\mathcal{T}} \rangle \right) \tag{3}$$

This objective encourages the real image-text pairs to come closer in the embedding space and pushes the fake pairs apart. Though this loss is on the labeled data, it helps the pseudo-labeling process for the unlabeled data.

Learning from Unlabeled Data: As explained earlier, we use model consensus to generate pseudo-labels for the unlabeled data, among which the confident ones satisfying the criterion in equation Eq. 1 are used for training the model as

Table 1: Data samples distribution for training, validation and testing for semi-supervised MFND task.

Dataset	Training		Validation	Testing
	Labeled (5%)	Unlabeled (95%)		
NewsCLIPpings [4]	3513	67555	7023	7263
GossipCop [28]	542	10302	2711	3388
PolitiFact [29]	12	225	59	74

follows:

$$\mathcal{L}_{CE}^{ul} = -\frac{1}{N^{ul}} \sum_{j=1}^{N^{ul}} \left(\tilde{y}_j^{\text{mc}} \log \hat{y}_j + (1 - \tilde{y}_j^{\text{mc}}) \log(1 - \hat{y}_j) \right) \tag{4}$$

Here, $\hat{y}_j$ be the model's predicted output for the j^{th} unlabel image-caption pair $(\mathcal{I}_j^{ul}, \mathcal{T}_j^{ul})$ and the total loss for model training is thus given by

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{CE}^{l} + \mathcal{L}_{CE}^{ul} + \lambda \mathcal{L}_{CC}^{l},$$

where λ is a hyperparameter used to balance the contribution of the contrastive loss. The complete training procedure is summarized in Algorithm 1. After training, the BLIP model need not be stored for inference, and only the trained CLIP model is used for inference as explained in Algorithm 2.

5 Experiments

Here, we discuss the datasets and implementation details, followed by the experimental results.

5.1 Dataset details

We train and evaluate the proposed approach on three well-known multi-modal FND datasets: NewsCLIPpings [4], GossipCop [28], and PolitiFact [29]. (i) NewsCLIPpings: Designed to address the evolving threat of misinformation from cheap fakes to sophisticated deep fakes, this dataset presents unmanipulated but contextually mismatched image-text pairs. (ii) GossipCop: Reflects the widespread nature of celebrity-related misinformation. (iii) PolitiFact: Created from news articles for fact-checking purposes. To the best of our knowledge, there are no existing works on SS-MFND. So we create the SS-MFND protocol by dividing the available datasets into a labeled and unlabeled part. Table 1 presents a detailed breakdown of the data samples distributed across the training, validation, and testing phases for each dataset for the semi-supervised MFND task. We also conducted experiments with different percentages of labeled and unlabeled data to test the robustness of the proposed approach.

5.2 Implementation Details

We use CLIP-ViT/B32 [8] for image and text encoding as the backbone and the BLIP captioning large model [19] for image captioning. The output of CLIP model features is followed by two fully connected layers for classification, which consists of batch normalization and dropout. Before the main training phase, both vision-language models (VLMs) are fine-tuned using the labeled data as a warm-up step. After the warm-up, we train the model for 40 epochs with a batch size of 64 across all datasets. The learning rate is set to 5×10^{-4} over the first 20 epochs and then follows a CosineAnnealing scheduler for the remaining epochs. The Adam optimizer, combined with this learning rate schedule, is employed for the training process. All experiments are conducted on NVIDIA RTX 2080 GPUs using the PyTorch library.

5.3 Baselines

Since there are no existing works which have addressed the SS-MFND task, we create our own strong baselines to compare the proposed framework. First, we consider the supervised training of the CLIP model with the available labeled data using cross-entropy loss, referred to as Sup@5%, which serves as the lower bound for our experiments. Additionally, we consider supervised training with 100% labeled data (also referred to as Sup@100%), representing the upper bound and the best possible performance achievable by utilizing unlabeled data and the given model. In addition, we also include three strong baselines using the state-of-the-art semi-supervised approaches proposed for uni-modal case, but adapted here for our task: FixMatch [16], FreeMatch* [18], and Adsh [17]. These methods leverage unlabeled data through different thresholding schemes to learn model representations. FixMatch employs a fixed threshold, while FreeMatch* uses learnable threshold parameters inspired by FreeMatch [18]. Since our approach focuses on obtaining optimal threshold parameters, we denote the version with learnable threshold parameters as FreeMatch*, and Adsh generates adaptive thresholds based on class dependencies. We have extensively fine-tuned these approaches for optimal threshold parameters and report their best performance. Inspired by MFND literature [4, 9, 11, 12], we report standard test accuracy as the performance metric for comparison.

5.4 Experimental Results

Table 2 reports the experimental results on NewsCLIPpings [4], GossipCop [28], and PolitiFact [29] datasets for the semi-supervised MFND task. For NewsCLIPpings, the CLIP model trained with only 5% labeled data (Sup@5%) achieved 65.57% accuracy. Utilizing the semi-supervised thresholding methods did not show any significant improvement, and in some cases even degraded the performance. This indicates that the thresholding methods on logit space proposed for image classification (unimodal case) may fail to capture the intricate relationship between real and fake image-text pairs. In contrast, the proposed CoVLM

Table 2: Experimental results on NewsCLIPpings, GossipCop and PolitiFact datasets.

Method	NewsCLIPpings	GossipCop	PolitiFact
Sup @ 5% (Lower Bound)	65.57%	51.91%	50.78%
FixMatch [16]	65.66%	73.70%	55.78%
FreeMatch* [18]	64.06%	74.21%	59.90%
Adsh [17]	64.78%	72.05%	50.78%
CoVLM (Ours)	**67.34%**	**76.42%**	**60.00%**
Sup @ 100% (Upper Bound)	70.02%	83.76%	72.50%

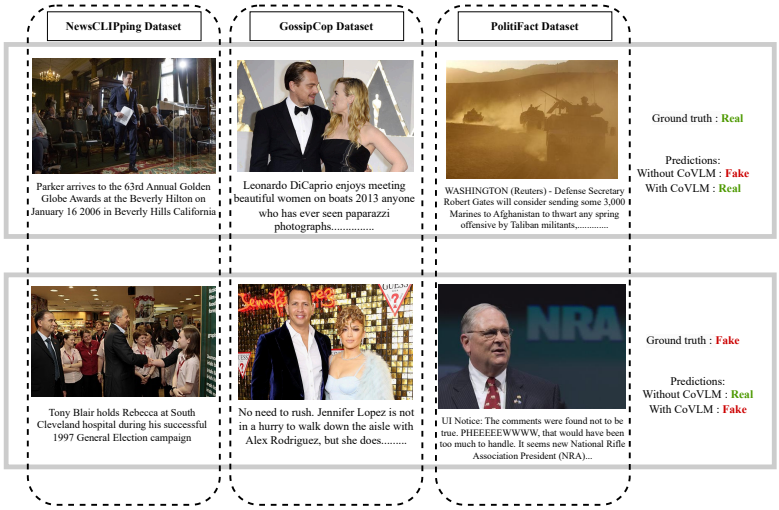


Fig. 6: The first row shows real image-text pairs, which were misclassified without CoVLM. After applying CoVLM, the model correctly predicts these pairs as real. The second row shows fake image-text pairs, which CoVLM accurately identifies as fake.

approach showed an improvement of 1.77% by effectively utilizing the unlabeled data.

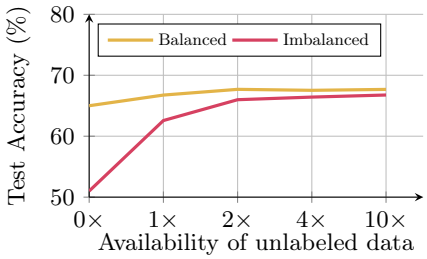
For the GossipCop dataset, the baseline accuracy with only 5% labeled data is 51.91%. However, the proposed CoVLM significantly improves this, achieving 76.42%. On the PolitiFact dataset, the baseline supervised accuracy is 50.78%. Here, CoVLM demonstrates substantial improvement, achieving 60.00%. Although there is still scope for improvement to reach the upper bound on these datasets, CoVLM significantly outperforms state-of-the-art semi-supervised approaches by a considerable margin.

Figure 6 shows a visual comparison highlighting the efficacy of the CoVLM method in identifying real and fake news across three distinct datasets: NewsCLIPpings, GossipCop, and PolitiFact. In the top row, the image-text pairs are from the Real class but were incorrectly predicted as Fake without CoVLM; with

Table 3: Experiment results on imbalanced NewsCLIPpings dataset.

Model	NewsCLIPpings
Sup @ 5%	51.89%
FixMatch [16]	63.84%
FreeMatch* [18]	54.36%
Adsh [17]	63.36%
CoVLM (Ours)	**66.76%**

Fig. 7: This graph illustrates the effect of amount of unlabeled Data on test accuracy on the NewsCLIPpings dataset.



CoVLM, they were correctly predicted as Real. In contrast, the bottom row displays instances of misinformation that, without CoVLM, were wrongly classified as true stories but were correctly identified as Fake with CoVLM.

6 Analysis and Ablation Study

In this section, we analyze the effects of data imbalance on the MFND task, the influence of the quantity of unlabeled data in the training procedure, and the impact of each loss component of the proposed CoVLM approach.

6.1 Impact of Data Imbalance in MFND

As mentioned earlier, in real-world, there exists a severe data imbalance in image-text pairs for the MFND task, making it more challenging. This is primarily because the number of real image-caption pairs is usually much higher than the fake ones. To simulate this challenging scenario, inspired from class-imbalance semi supervised learning [17], we synthetically imbalance the NewsCLIPpings dataset with a 9:1 ratio, meaning that out of every 10 samples, 9 will be real and 1 will be fake in both labeled and unlabeled data. This imbalance makes the task more challenging and can cause the model to become biased towards predicting every sample as real, unless suitable measures are taken to handle this imbalance.

Table 3 shows the experimental results for this imbalanced SS-MFND protocol. We observe that for this scenario, methods like FixMatch and Adsh helps to improve the baseline performance to a great extent. But the proposed CoVLM framework significantly outperforms all the other methods, with results close to that of the balanced case.

6.2 Impact of Amount of Unlabeled Data

In real-world, the amount of unlabeled data available for training the model can vary. To analyze the proposed framework under these conditions, we evaluate its

Table 4: Ablation study on NewsCLIPpings Dataset illustrating the importance of each component in the loss function.

$\mathcal{L}_{CE}^{l}$	$\mathcal{L}_{CC}^{l}$	$\mathcal{L}_{CE}^{ul}$	Balanced	Imbalanced
✓	×	×	65.77%	51.89%
✓	✓	×	66.57%	63.96%
✓	✓	✓	**67.67%**	**66.76%**

performance by varying the amounts of unlabeled data relative to the labeled data. The amount of labeled data (5%) is fixed for all these cases, only the unlabeled data is varied. Specifically, we use the number of unlabeled data as $0\times, 1\times, 2\times, 4\times,$ and $10\times$ the number of labeled data. From Figure 7, we observe that the model's performance saturates at approximately $4\times$ the amount of labeled data in both balanced and imbalanced cases.

6.3 Ablation Study

The proposed framework is trained using three loss components. To demonstrate the importance of each proposed component, we conducted an ablation study on the NewsCLIPpings dataset. Table 4 presents the results for both balanced and imbalanced cases. The first row represents the model trained only with the labeled cross-entropy loss ($\mathcal{L}_{CE}^{l}$), which essentially fine-tunes CLIP without additional guidance. Adding the contrastive clustering loss ($\mathcal{L}_{CC}^{l}$) significantly enhances the performance by enforcing better separation between real and fake pairs. Finally, incorporating the unlabeled data loss ($\mathcal{L}_{CE}^{ul}$) with robust pseudo-labels generated from the consensus of CLIP and BLIP further boosts performance, showcasing the benefits of CoVLM for semi-supervised MFND.

7 Conclusion

In this paper, we introduced CoVLM, a novel framework for semi-supervised multi-modal fake news detection, designed to operate effectively with limited labeled data and a substantial amount of unlabeled data. By leveraging the consensus between two vision-language models, CLIP and BLIP, CoVLM generates robust pseudo-labels that capture the intricate relationships between images and text. Our extensive experiments on benchmark datasets such as NewsCLIPpings, GossipCop, and PolitiFact demonstrate the effectiveness of CoVLM. Moreover, CoVLM effectively handles data imbalance, maintaining its performance even when the datasets are imbalanced to reflect real-world conditions. This work offers a significant advancement in multi-modal fake news detection, providing a practical solution for leveraging both labeled and unlabeled data.

Acknowledgement This work is partly supported through a research grant from SERB, Department of Science and Technology, Govt. of India (SPF/2021/000118).

References

1. Mohammed Rasheed Omar and Adnan Mohsin Abdulazeez. Fake news in social network: A comprehensive review. *Indonesian Journal of Computer Science*, 2024.
2. Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.
3. Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.
4. Grace Luo, Trevor Darrell, and Anna Rohrbach. Newsclippings: Automatic generation of out-of-context multimodal media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.
5. Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.
6. Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8):2939–2970, 2022.
7. Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. Self-supervised multimodal learning: A survey. *arXiv preprint arXiv:2304.01008*, 2023.
8. Alec Radford et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.
9. Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
10. Xin Yuan, Jie Guo, Weidong Qiu, Zheng Huang, and Shujun Li. Support or refute: Analyzing the stance of evidence to detect out-of-context mis- and disinformation. *arXiv preprint arXiv:2311.16496*, 2023.
11. Fatma Shalabi, Huy H. Nguyen, Hichem Felouat, Ching-Chun Chang, and Isao Echizen. Image-text out-of-context detection using synthetic multimodal misinformation. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, 2024.
12. Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
13. Xuannan Liu, Zekun Li, Peipei Li, Shuhan Xia, Xing Cui, Linzhi Huang, Huaibo Huang, Weihong Deng, and Zhaofeng He. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. *arXiv preprint arXiv:2406.08772*, 2024.
14. Yangming Zhou, Qichao Ying, Zhenxing Qian, Sheng Li, and Xinpeng Zhang. Multimodal fake news detection via clip-guided learning. *arXiv preprint arXiv:2205.14304*, 2022.
15. Ye Jiang, Xiaomin Yu, Yimin Wang, Xiaoman Xu, Xingyi Song, and Diana Maynard. Sample: Similarity-aware multimodal prompt learning for fake news detection. *arXiv preprint arXiv:2304.04187*, 2023.

16. Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.
17. Lan-Zhe Guo and Yu-Feng Li. Class-imbalanced semi-supervised learning with adaptive thresholding. *Proceedings of the 39th International Conference on Machine Learning*, 2022.
18. Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Zhen Wu, and Jindong Wang. Freematch: Self-adaptive thresholding for semi-supervised learning. *ArXiv*, abs/2205.07246, 2022.
19. Jiahui Li, Li Dong, Sheng Wang, and Cees G. M. Snoek. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.
20. Minsung Hyun, Jisoo Jeong, and Nojun Kwak. Class-imbalanced semi-supervised learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. Suppressed Consistency Loss (SCL) method.
21. Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
22. Youngtaek Oh, Dong-Jin Kim, and In So Kweon. Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
23. Zhen Jiang, Lingyun Zhao, Yu Lu, Yongzhao Zhan, and Qirong Mao. A semi-supervised resampling method for class-imbalanced learning. *Journal of Machine Learning Research*, 2021.
24. Elaheh Arabmakki. A reduced labeled samples (rls) framework for classification of imbalanced concept-drifting streaming data. *ThinkIR: The University of Louisville's Institutional Repository*, 2016.
25. Lefan Zhang, Zhang-Hao Tian, Wujun Zhou, and Wei Wang. Learning from long-tailed noisy data with sample selection and balanced loss. *Neural Networks*, 2020.
26. Y. Fan, D. Dai, and B. Schiele. CoSSL: Co-learning of representation and classifier for imbalanced semi-supervised learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
27. H. Chen, Y. Fan, Y. Wang, J. Wang, B. Schiele, X. Xie, M. Savvides, and B. Raj. An embarrassingly simple baseline for imbalanced semi-supervised learning. *IEEE Access*, 2021. Introduces SimiS method.
28. Gossipcop dataset. Available at: `www.gossipcop.com`.
29. Gossipcop dataset. Available at: `www.politifact.com`.
30. Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, pages 141–161, 2020.
31. Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14185–14193, 2021.
32. Bing Han, Xiaoguang Han, Hua Zhang, Jingzhi Li, and Xiaochun Cao. Fighting fake news: two stream network for deepfake detection via learnable srm. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3):320–331, 2021.

33. Peiguang Li, Xian Sun, Hongfeng Yu, Yu Tian, Fanglong Yao, and Guangluan Xu. Entity-oriented multi-modal alignment and fusion network for fake news detection. *IEEE Transactions on Multimedia*, 24:3455–3468, 2021.
34. Nadia K Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1):1–4, 2015.
35. Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.
36. Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347, 2021.
37. Xinyi Zhou, Jindi Wu, and Reza Zafarani. Safe: Similarity-aware multi-modal fake news detection. *arXiv preprint arXiv:2003.04981*, 2020.
38. Priyanshi Virendra Shah. Multimodal fake news detection using a cultural algorithm with situational and normative knowledge. *arXiv preprint arXiv:2007.07045*, 2020.
39. Anastasia Giachanou, Guobiao Zhang, and Paolo Rosso. Multimodal fake news detection with textual, visual and semantic information. In *Springer*, 2020.
40. Isabel Segura-Bedmar and Santiago Alonso Bartolomé. Multimodal fake news detection. *MDPI Information*, 13(6):284, 2022.
41. Ashima Yadav, Shivani Gaba, Haneef Khan, Ishan Budhiraja, Akansha Singh, and Krishan Kant Singh. Etma: Efficient transformer based multilevel attention framework for multimodal fake news detection. *arXiv preprint arXiv:2206.07331*, 2022.
42. Kamonashish Saha. Debertnext: A multimodal fake news detection framework. *University of Windsor Leddy Library*, 2023.
43. Anonymous. Tt-blip: Enhancing fake news detection using blip and tri-transformer. *arXiv preprint arXiv:2403.12481*, 2024.
44. Arka Ujjal Dey, Artemis Llabrés, Ernest Valveny, and Dimosthenis Karatzas. Retrieval augmented verification: Unveiling disinformation with structured representations for zero-shot real-time evidence-guided fact-checking of multi-modal social media posts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2024.
45. Zacharias Chrysidis, Stefanos-Iordanis Papadopoulos, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. Credible, unreliable or leaked?: Evidence verification for enhanced automated fact-checking. *arXiv preprint arXiv:2404.18971*, 2024.
46. David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2019.
47. Qinyi Deng, Yong Guo, Zhibang Yang, Haolin Pan, and Jian Chen. Boosting semi-supervised learning with contrastive complementary labeling. *arXiv preprint arXiv:2212.06643*, 2022.
48. Jongjin Park, Sukmin Yun, Jongheon Jeong, and Jinwoo Shin. Opencos: Contrastive semi-supervised learning for handling open-set unlabeled data. *arXiv preprint arXiv:2107.08943*, 2021.
49. Jiwon Kim, Youngjo Min, Daehwan Kim, Gyuseong Lee, Junyoung Seo, Kwangrok Ryoo, and Seungryong Kim. Conmatch: Semi-supervised learning with confidence-guided consistency regularization. *arXiv preprint arXiv:2208.08631*, 2022.

50. Erik Wallin, Lennart Svensson, Fredrik Kahl, and Lars Hammarstrand. Doublematch: Improving semi-supervised learning with self-supervision. *arXiv preprint arXiv:2205.05575*, 2022.
51. Khanh-Binh Nguyen and Joon-Sung Yang. Boosting semi-supervised learning by bridging high and low-confidence predictions. *arXiv preprint arXiv:2308.07509*, 2023.
52. Hieu Pham, Zihang Dai, Qizhe Xie, Minh-Thang Luong, and Quoc V. Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
53. Yue Duan, Zhen Zhao, Lei Qi, Luping Zhou, Lei Wang, Yinghuan Shi, and Yang Gao. Mutexmatch: Semi-supervised learning with mutex-based consistency regularization. *arXiv preprint arXiv:2203.12359*, 2022.
54. Junnan Li, Caiming Xiong, and Steven Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
55. David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *Advances in Neural Information Processing Systems*, 2019.
56. Kamal Nigam, Andrew McCallum, and Tom Mitchell. *Semi-Supervised Text Classification Using EM*, chapter Semi-Supervised Text Classification Using EM. 2006.
57. Shweta Dharmadhikari, Maya Ingle, et al. Analysis of semi supervised learning methods towards multi label text classification. *International Journal of Computer Applications*, 2012.
58. D. Barman and N. Chowdhury. A novel semi supervised approach for text classification. *Int. j. inf. tecnol.*, 12:1147–1157, 2020.
59. Alexander Hanbo Li and Abhinav Sethy. Semi-supervised learning for text classification by layer partitioning. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6164–6168, 2020.
60. Yang Li, Ying Lv, Suge Wang, Jiye Liang, Juanzi Li, and Xiaoli Li. Cooperative hybrid semi-supervised learning for text sentiment classification. *Symmetry*, 11(2):133, 2019.
61. Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification, 2021. arXiv:1605.07725 [stat.ML].
62. J.M. Duarte and L. Berton. A review of semi-supervised learning for text classification. *Artif Intell Rev*, 56:9401–9469, 2023.
63. Edson Takashi Matsubara, Maria Carolina Monard, and Gustavo E. A. P. A. Batista. Multi-view semi-supervised learning: An approach to obtain different views from text datasets. *Institute of Mathematics and Computer Science – ICMC*, 2024.